Assessing Fair Policing in Austin, TX

Team FunkyStats

4/24/2021

Introduction

Our presentation consists of three parts:

- Exploratory data analysis to get a big picture of policing in Austin:
  - Benchmark Test
  - Outcome Test
  - Veil of Darkness Test
- Various modeling strategies to assess the severity of racial disparities:
  - Logistic Regression
  - Bayesian Hierarchical Model
- Propose a measure of fairness
  - based on the differences in the posterior median hit rate among individual police officers

Available Data

- Stanford Open Policing Project data (2006.01.01 - 2016.06.30, 463,944 stops): stops time, the driver race, searched or frisked, contraband discovered etc.
  - Merits: contain driver race
  - Drawbacks: missing time and location information
- APD Racial Profiling data (2019, 79,693 stops):
  - Merits: contain time, location, and officer race
  - Drawback: missing driver race
- US census demographic data (2012-2017 5-year, 2019 ACS)
- APD Racial Profiling Report

Number of stops

We note that the distribution of stops per officer has an extremely long tail.

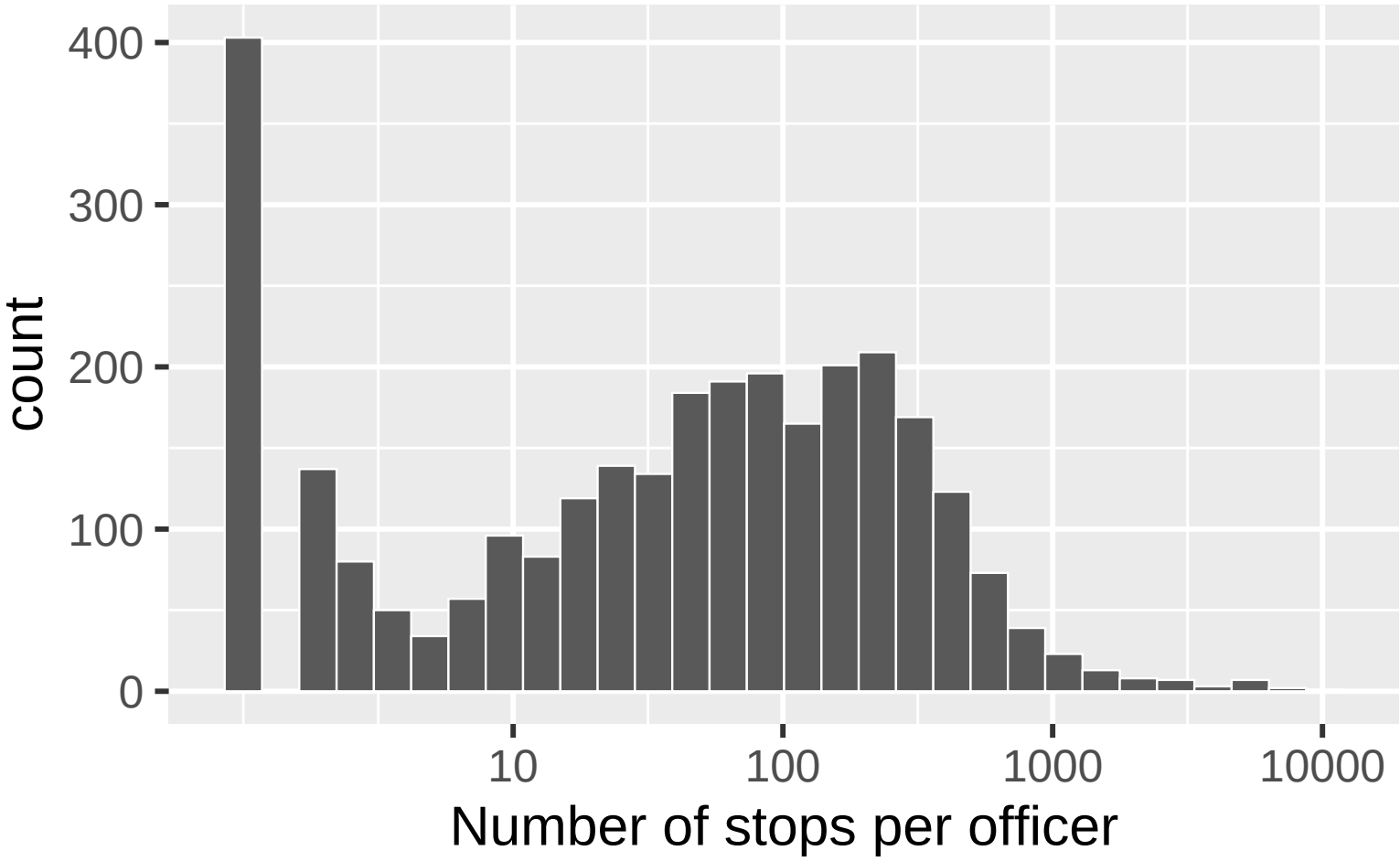


Figure 1: Distribution of stops by unique officer ID

Exploratory Analysis

Examining the count of stops by race during 2006-2015:

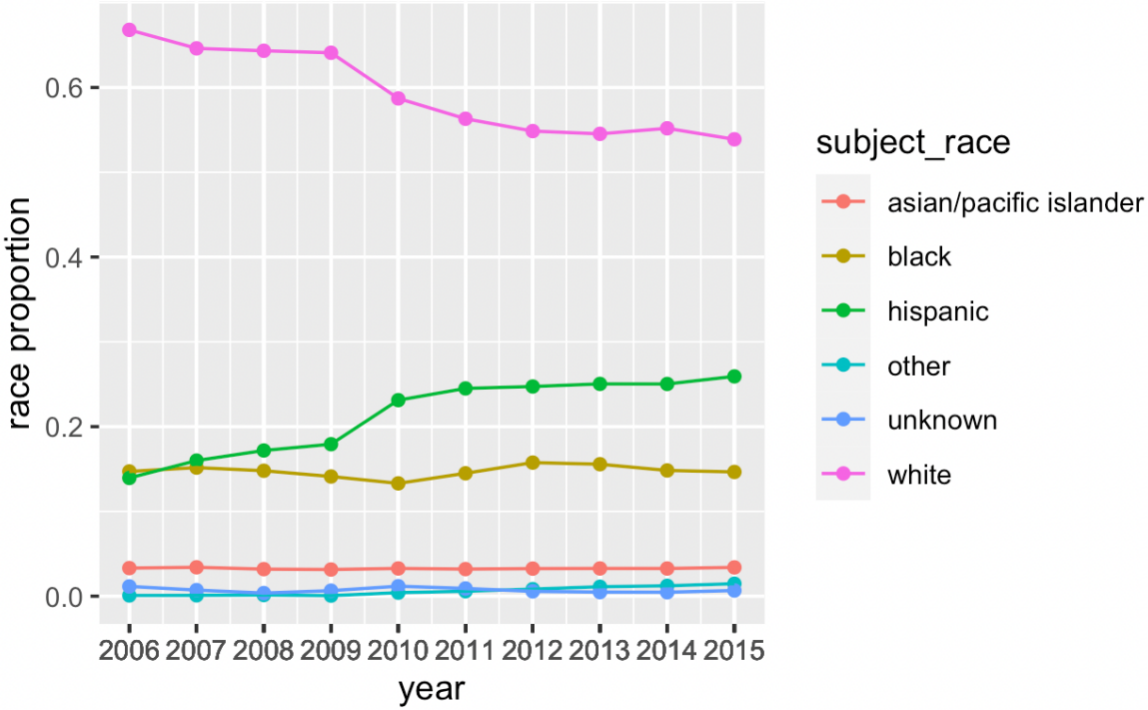
- More than half of the stops were of white drivers; the number of stops for white drivers is roughly four times that of Black drivers.
- The white population in Austin (445,269) is almost 7 times larger than the Black population (66,724)

Driver Race	Counts	Proportion
Asian/Pacific	11658	0.033
Black	52381	0.147
Hispanic	765707	0.215
Other	2105	0.006
Unknown	2622	0.007
White	211588	0.593

Exploratory Analysis

Examine race proportion in each year:

- Annual trends are very different by race
- Fewer white drivers stopped especially after 2009
- An increasing trend of Hispanic and black drivers being stopped



Benchmark Test

$$\text{Stop Rate}_i = \frac{\text{Number of Stops for Race } i}{\text{Population of Race } i}$$

$$\text{Search Rate}_i = \frac{\text{Number of Stopped People Who Were Searched for Race } i}{\text{Number of Stops for Race } i}$$

$$\text{Frisk Rate}_i = \frac{\text{Number of Stopped People Who Were Frisked for Race } i}{\text{Number of Stops for Race } i}$$

Driver Race	Counts	Population	Proportion	Stop Rate	Search Rate	Frisk Rate	Hit Rate
asian/pacific	11658	63752	0.033	0.183	0.015	0.011	0.188
black	52381	66724	0.147	0.785	0.092	0.039	0.254
hispanic	765707	316709	0.215	0.242	0.086	0.044	0.323
white	211588	445269	0.593	0.475	0.031	0.021	0.318

Benchmark Test Caveats

- Insufficient evidence of discriminative policing
- We seek the true distribution of the drivers violating the law
- Check if different race groups are disproportionately stopped

Outcome Test

- Define a successful search as one that uncovers contraband
- Hit rate is the proportion of searches that are successful
  - If racial groups have different hit rates, it can be taken as evidence of discriminative policing

$$\text{Hit Rate}_i = \frac{\text{Number of Contraband Uncovered for Race } i}{\text{Number of Searched People for Race } i}$$

Driver Race	Counts	Population	Proportion	Stop Rate	Search Rate	Frisk Rate	Hit Rate
asian/pacific	11658	63752	0.033	0.183	0.015	0.011	0.188
black	52381	66724	0.147	0.785	0.092	0.039	0.254
hispanic	765707	316709	0.215	0.242	0.086	0.044	0.323
white	211588	445269	0.593	0.475	0.031	0.021	0.318

Outcome Test Caveats

- Only outcomes available: Although the outcome test is simple and intuitive, the actual threshold for searching someone is not observed.
- Infra-marginality problem and subgroup validity problem

Veil of Darkness Test

- Hypothesis: officers who are engaged in racial profiling are less likely to be able to identify a driver's race after dark than during daylight
- Under this hypothesis, if stops made after dark had smaller proportion of black drivers stopped than stops made during daylight, it could be evidence of racial profiling.
- Two key elements: Driver race & Stop time
- Alternative: measure the racial population in different areas through zip codes. If the number of the stops made during daytime and nighttime in areas with a larger Black population than white population show substantially different patterns than stops in areas with larger white populations, it could be evidence of racial profiling.

Veil of Darkness Test

In order to accurately distinguish the daytime and nighttime, we compute the daily sunset and dusk time for Austin in 2019.

- Earliest sunset in 2019 was at around 17:32 in early December and it goes fully dark in 26 minutes
- Latest sunset time was around 20:38 late June and it was fully dark after 28 minutes.

Date	Sunset	Dusk
2019-12-02	17:31:42	17:57:48
2019-06-30	20:37:58	21:05:27

For simplicity, here we consider only the black and the white population.

- Daytime Stop: stops happening before sunset
- Nighttime Stop: stop happening after dusk
- Majority-Black Area (MBA): areas with a larger Black population than white population

Veil of Darkness Test

	Day	Night
MBA	124	126
MWA	2937	2216

- Assume two rows as independent binomial samples
- Of $n_1 = 250$ recorded stops in majority-Black area, 124 stops happened during the daytime, a proportion of $p_1 = 124/250 = 0.496$
- Of $n_2 = 5153$ recorded stops in majority-white area, 2937 stops happened during the daytime, a proportion of $p_2 = 2937/5153 = 0.570$
- The sample difference of proportions is 0.074
- We obtain Fisher's exact test for testing null hypothesis of independence of the two rows with p value of 0.02, indicating the strong evidence that the police are not equally likely practicing during day and night to different racial groups.

Hit Rate and Causal Issues

- Unmeasured confounders: Crime rates are known to be correlated with income and demographic factors
  - More officers patrolled areas with higher crime rates
  - Neighborhoods with higher minority populations is expected to see more minority traffic stops
- Although this still exposes problems in Austin, it could be interpreted as a problem of economic segregation, not traffic fairness.
- To overcome this problem, we propose looking at the hit rate, with more details to follow:
  - Given that an officer has decided to search or frisk a subject, the probability of finding contraband should be equal among all races, regardless of the neighborhood in which the search was conducted (equal evidence thresholds).
  - Using the hit rate does not eliminate all unmeasured confounders, but it helps mitigate the problem.

Logistic Regression

Logistic Regression for Frisk Rate

We want to seek given a person is stopped, what factors may impact the likelihood of that person being frisked? We fit a logistic regression model with `frisk` as the dependent variable of `race`, `age`, and `sex`.

$$\text{Logit}[\text{P}(\text{Being Frisked})] = \beta_0 + \beta_1 \text{Race} + \beta_2 \text{Age} + \beta_3 \text{Sex}$$

Logistic Regression for Contraband found

We want to investigate how likely contraband items are found when searching is performed. We argue that if racial bias does not exist, the hit rate should be equal for all races.

$$\text{Logit}[\text{P}(\text{Contraband found})] = \beta_0 + \beta_1 \text{Race} + \beta_2 \text{Age} + \beta_3 \text{Sex}$$

We also break down contraband found into three categories: `Drugs`, `Weapons`, and `Others`, and fit a logistic regression model with `Race`.

Logistic Regression

- Black and Hispanic drivers are more likely to be frisked than white drivers
  - The estimated odd of a Black driver being frisked during a stop is 2.19 time that of white drivers. For Hispanic drivers, the odds ratio is 1.80.
  - Asian people are the least likely to be frisked.
- Contraband items are more likely to be found when searching Hispanic and Black drivers
  - White people are more likely to be found with weapons
  - Black and Hispanic people are more likely to be found with contraband items that are neither drugs or weapons

Bayesian Modeling

- We use hierarchical models to investigate hit rates at the officer level. In the following analysis, we define the hit rate as the proportion of times an officer finds contraband given that a frisk has been performed.
- The hit rate can be thought of as a proxy for "evidence" when an officer decides whether to conduct a search or a frisk.
- We are interested in the hit rates for each subpopulation (white, Black, Hispanic) for each officer.
  - a lower hit rate for a particular subpopulation could imply a lower threshold of evidence when policing that subpopulation.
- Simplifying restrictions:
  - Consider only white, Black, and Hispanic subpopulations.
  - Consider only officers with 18 or more stops (roughly the 90th percentile and above).

Visualizing individual officer hit rates

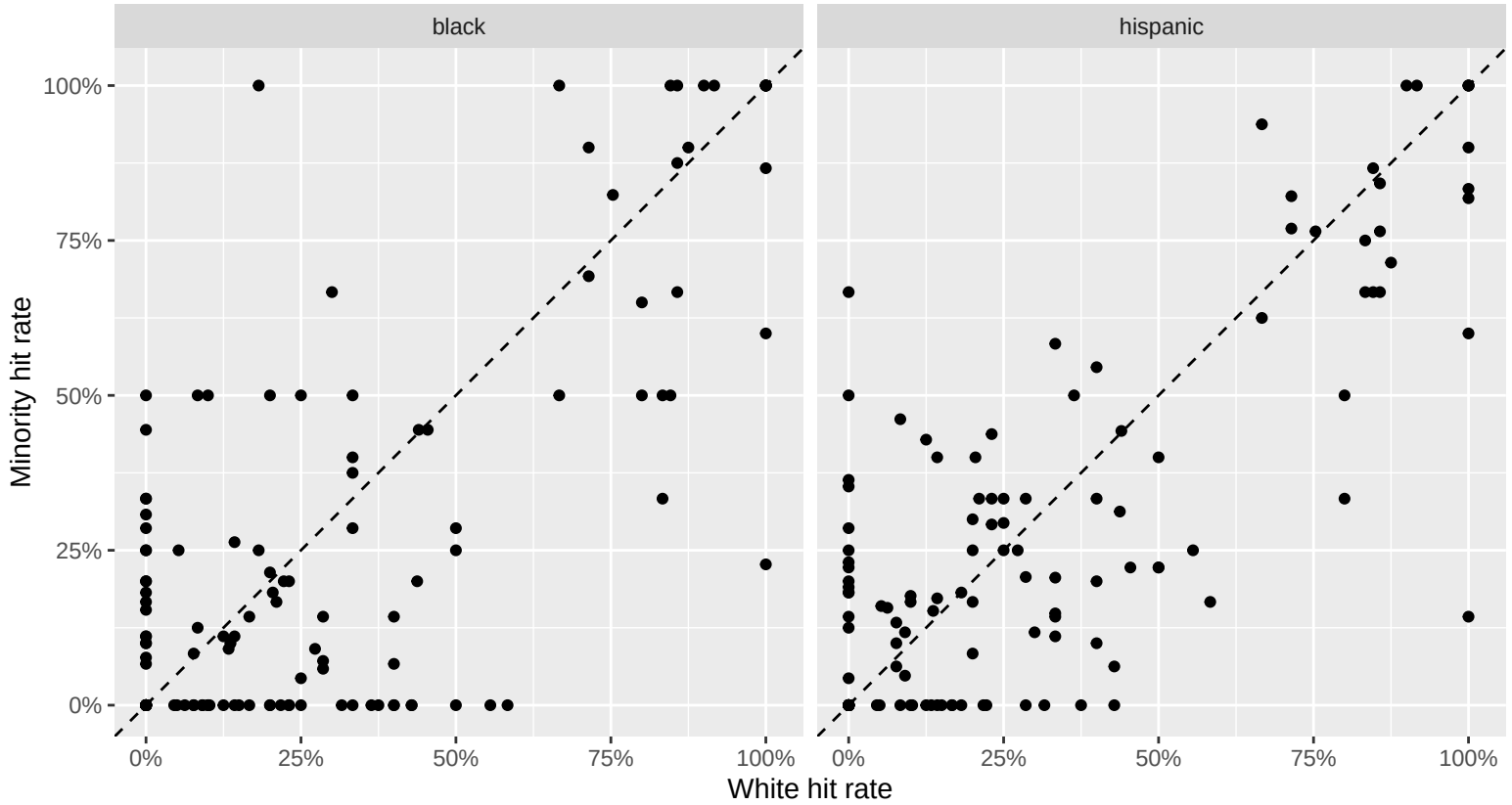


Figure 2: Hit rates for individual officers.

- We can think of individual officers as belonging to a population of officers; we want to model both the hit rates of the individual officers and the variation of this population.
- Because we are interested in hit rates for three different subpopulations, we fit three separate hierarchical models using different subsets of the data.
    - For instance, we model the hit rate for Black subjects by directly subsetting the data to include only stops of Black subjects.

Why hierarchical models?

- Hierarchical models allow for *partial pooling*, by which individual hit rates are biased towards the population average by an amount determined by the estimates of the population parameters and the data available for each officer.
- Why is partial pooling desirable in this case?
  - Observed hit rates at the boundaries will have posteriors containing more reasonable values.
    - No pooling will almost certainly overestimate for officers with perfect or near-perfect hit rates. Likewise, it underestimates for officers with zero or near-zero hit rates.
  - Complete pooling (equal hit rates for all officers) is unrealistic, as individual officers may have different overall thresholds of evidence and may be more or less experienced.
  - We have a different number of stops for each officer

Specifically, let θ_{jr} be the hit rate for for officer j and race r, y_{jr} be the number of hits, and K_{jr} the number of frisks. In the following, because we fit separate models, we assume for example $r = \textit{black}$ and drop the r subscript. Assuming each officer's searches are independent Bernoulli trials

$$p(y_j|\theta_j) = \text{Binomial}(y_j|K_j, \theta_j)$$

We reparametrize the model in terms of the log-odds, α:

$$\alpha_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j}$$

We use a weakly informative Normal prior for α_j centered at -1.3 and with standard deviation 1 ($\text{logit}^{-1}(\alpha_j) \approx .2$). The model is therefore

$$p(y_j|K_j, \alpha) = \text{Binomial}(y_j|K_j, \text{logit}^{-1}(\alpha_j))$$

We proceed using `rstanarm::stan_glmer` and the default prior on the covariance matrix.

- The result includes a posterior for each officer; we may transform from the log-odds back to hit rate to to obtain a posterior for the the hit rate for each officer.
- We model each race separately, and so obtain three posteriors for each officer.

Table 1: Posterior intervals for several officers for three races. From left to right: white, Black, Hispanic

ID	W 2.5%	W 50%	W 97.5%	B 2.5%	B 50%	B 97.5%	H 2.5%	H 50%	H 97.5%
01db7098a7	0.003	0.055	0.317	0.021	0.198	0.617	0.065	0.214	0.450
020579eaad	0.021	0.095	0.248	0.002	0.033	0.203	0.047	0.162	0.359
02b0803fe3	0.003	0.056	0.324	0.111	0.242	0.420	0.037	0.174	0.423
0329f48f95	0.128	0.523	0.900	0.057	0.521	0.959	0.665	0.874	0.974
068ff01d47	0.054	0.188	0.415	0.032	0.332	0.833	0.082	0.268	0.556

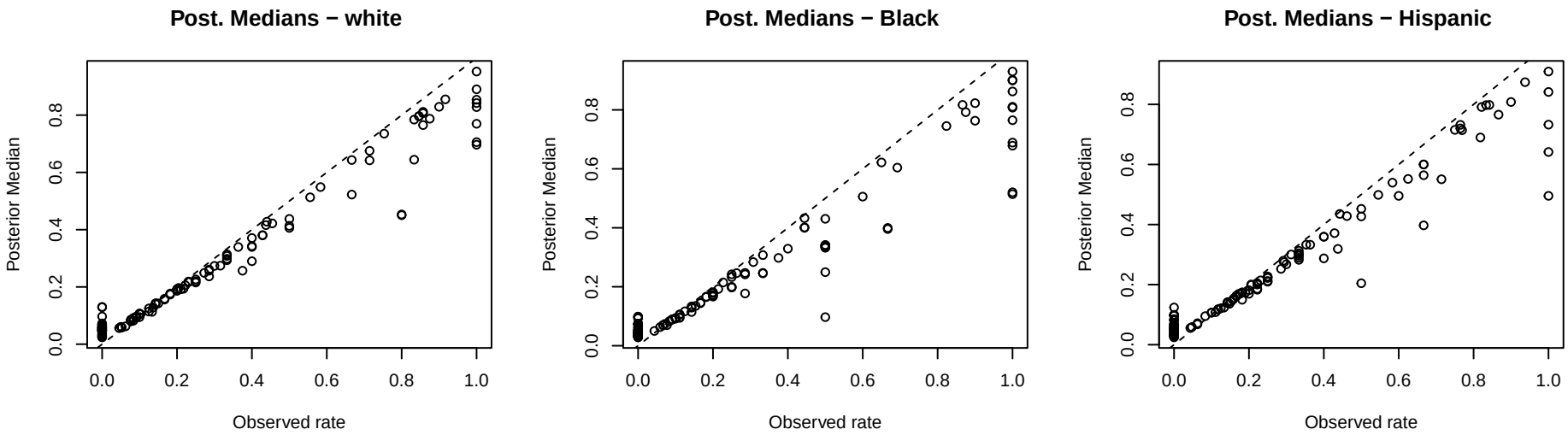


Figure 3: Partial pooling.

Operationalizing Fairness

- Because part of this project is to "operationalize" fairness, we devised a measure by which the above posteriors can be converted into a rough "fairness score."
- Supposing that an officer uses the same evidence threshold when deciding whether to frisk a subject regardless of the race of the subject, we would expect that officer to have roughly equal hit rates for all three subpopulations.
- We reason that such an officer should have posterior medians that are close to each other for the three subpopulations.

- So, one can calculate a simple sum of squares statistic for each officer. Specifically, letting m_{jr} be the posterior median for officer j and race r, the sum of squares statistic S_j is

$$S_j = \sum_r (m_{jr} - \bar{m}_j)^2$$

  where $\bar{m}_j$ is the average of the three medians.
- Of course, this measure disregards all other information that could be gleaned from the posterior!
  - An alternative might calculate the overlap between the posterior densities. However, we think this measure is relatively easy to understand and implement.

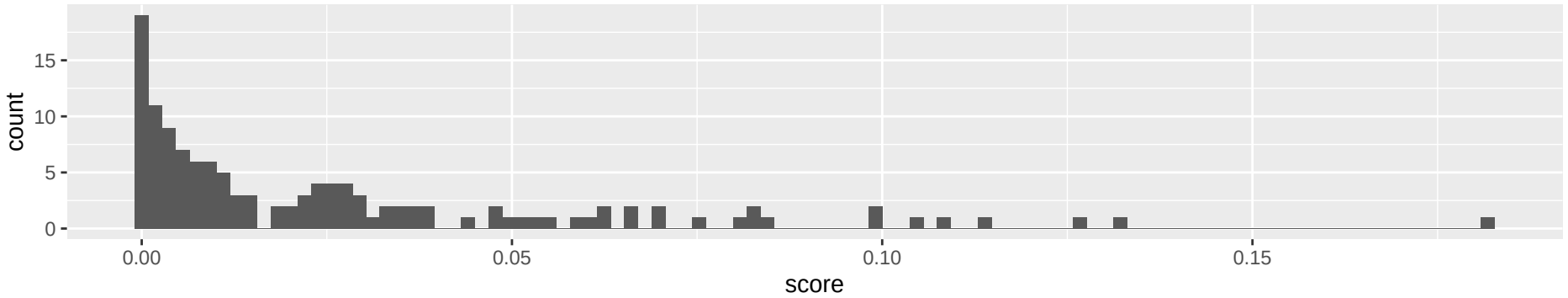


Figure 4: Fairness scores for the officers under consideration. Lower scores indicate hit rates are more similar.

EX: 5 Highest Scores

Table 2: Highest 5 scores.

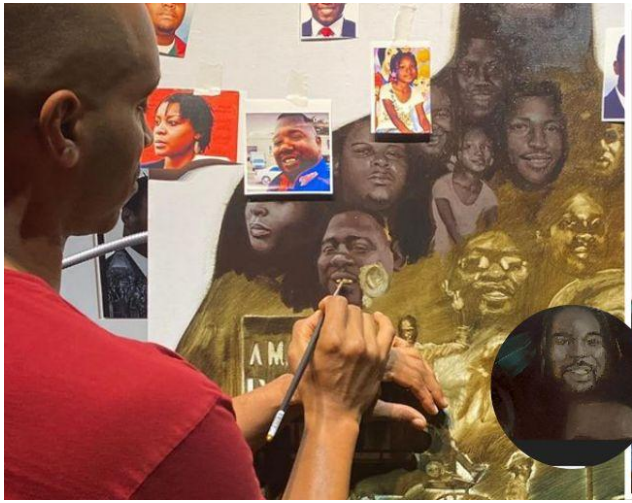
ID	W. Obs	B. Obs	H. Obs	W. Post.	B. Post.	H. Post.	W. Count	B. Count	H. Count
1504c3bc16	1.000	0.227	0.143	0.697	0.215	0.142	2	22	7
3392a495a3	0.400	0.000	0.545	0.371	0.046	0.499	10	5	11
50f70c6ecb	0.583	0.000	0.167	0.549	0.064	0.162	12	3	18
bab7c2acaf	0.833	0.333	0.750	0.644	0.247	0.715	7	3	20
dd9c1003d5	0.556	0.000	0.250	0.513	0.043	0.210	9	6	4

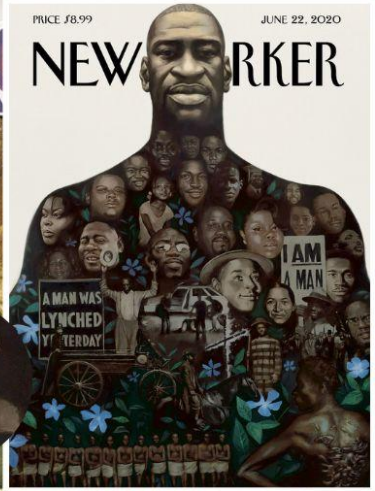
In the above table, "Obs." columns contain observed hit rates; "Post." columns contain posterior median hit rates; "Count" columns contain the number of instances.

Conclusion

- Three Tests (benchmark test, outcome test and veil of darkness test):
  - Evaluate the fairness of traffic stops
  - Confirm racial disparity in policing exists and is present in different scales
- Frequentist Modeling:
  - Explore the causal confounding issues through logistic regression
  - Conclude black and Hispanic people are more likely to be frisked and found with contraband items that are neither drugs or weapons
- Bayesian Modeling:
  - Investigate the hit rate via Bayesian hierarchical modeling
  - Obtain posteriors for the hit rate for each officer in a subset of the data
  - Devised a "fairness score" from the posteriors medians, a tool we believe could be used to identify officers with racially disparate patterns of traffic stops

Confront Systemic Racism





- Left: "Say Their Names" by Kadir Nelson (The New Yorker 2020)
- Right: Derek Chauvin guilty of George Floyd's murder, Apr 20, 2021

This is not justice.
This is accountability.
Chauvin is where we start.
The whole system is next.

Thank you for your attention!

UPSTAT
Upstate New York Statistics Conference

FunkyStats: David Skrill, Qiuyi Wu, Cuong Pham (from left to right)

References

Grogger, Jeffrey, and Greg Ridgeway. 2006. *Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness*. Santa Monica, CA: American Statistical Association.

"Hierarchical Partial Pooling for Repeated Binary Trials." n.d. https://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736–45. https://doi.org/10.1038/s41562-020-0858-1.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. 2017. "THE Problem of Infra-Marginality in Outcome Tests for Discrimination." *The Annals of Applied Statistics* 11 (3): 1193–1216. http://www.jstor.org/stable/26362224.